Multi-level computational methods for interdisciplinary research in the HathiTrust Digital Library

Murdock, Jaimie; Allen, Colin; Börner , Katy ; Light, Robert ; McAlister, Simon; Ravenscroft, Andrew

Published in:
PLoS ONE

DOI:
10.1371/journal.pone.0184188

Publication date:
2017

PLOS ONE

RESEARCH ARTICLE

Multi-level computational methods for interdisciplinary research in the HathiTrust Digital Library

Jaimie Murdock[1,2], Colin Allen[1,3,4‡*], Katy Börner[1,2,5,6‡], Robert Light[2], Simon McAlister[7], Andrew Ravenscroft[7‡], Robert Rose[1,8], Doori Rose[1], Jun Otsuka[9], David Bourget[10‡], John Lawrence[11], Chris Reed[11‡]

1 Program in Cognitive Science, Indiana University, Bloomington, IN, United States of America, **2** School of Informatics and Computing, Indiana University, Bloomington, IN, United States of America, **3** Department of History & Philosophy of Science & Medicine, Indiana University, Bloomington, IN, United States of America, **4** Department of History & Philosophy of Science, University of Pittsburgh, Pittsburgh, PA, United States of America, **5** Indiana University Network Science Institute (IUNI), Bloomington, IN, United States of America, **6** User-Centered Social Media, Department of Computer Science and Applied Cognitive Science, University of Duisburg-Essen, Duisburg, Germany, **7** International Centre for Public Pedagogy (ICPuP), Cass School of Education & Communities, University of East London, London, United Kingdom, **8** Department of Mathematics, Indiana University, Bloomington, IN, United States of America, **9** Department of Philosophy, Kyoto University, Kyoto, Japan, **10** Department of Philosophy, University of Western Ontario, London, Ontario, Canada, **11** Centre for Argument Technology, University of Dundee, Dundee, United Kingdom

‡These authors were the project leaders; see "Author's Contributions" for details of all contributions.
* prof.colin.allen@gmail.com

Citation: Murdock J, Allen C, Börner K, Light R, McAlister S, Ravenscroft A, et al. (2017) Multi-level computational methods for interdisciplinary research in the HathiTrust Digital Library. PLoS ONE 12(9): e0184188. https://doi.org/10.1371/journal.pone.0184188

Editor: Boris Podobnik, University of Rijeka, CROATIA

Received: February 9, 2017

Accepted: July 10, 2017

Published: September 18, 2017

Data Availability Statement: Corpus and model files are accessible at the IU ScholarWorks repository through the following URL: https://scholarworks.iu.edu/dspace/handle/2022/21636. The LoC-UCSD crosswalk is available on GitHub at https://github.com/inpho/loc-ucsd. These data and models are available without restrictions. Additional access to the raw corpus text is available at the HathiTrust Research Center (HTRC) via the Research Portal at http://analytics.hathitrust.org/, subject to restrictions implied by HathiTrust (see https://www.hathitrust.org/help_copyright).

Abstract

We show how faceted search using a combination of traditional classification systems and mixed-membership topic models can go beyond keyword search to inform resource discovery, hypothesis formulation, and argument extraction for interdisciplinary research. Our test domain is the history and philosophy of scientific work on animal mind and cognition. The methods can be generalized to other research areas and ultimately support a system for semi-automatic identification of argument structures. We provide a case study for the application of the methods to the problem of identifying and extracting arguments about anthropomorphism during a critical period in the development of comparative psychology. We show how a combination of classification systems and mixed-membership models trained over large digital libraries can inform resource discovery in this domain. Through a novel approach of "drill-down" topic modeling—simultaneously reducing both the size of the corpus and the unit of analysis—we are able to reduce a large collection of fulltext volumes to a much smaller set of pages within six focal volumes containing arguments of interest to historians and philosophers of comparative psychology. The volumes identified in this way did not appear among the first ten results of the keyword search in the HathiTrust digital library and the pages bear the kind of "close reading" needed to generate original interpretations that is the heart of scholarly work in the humanities. Zooming back out, we provide a way to place the books onto a map of science originally constructed from very different data and for different purposes. The multilevel approach advances understanding of the intellectual and societal contexts in which writings are interpreted.

Funding: This work was funded by the National Endowment for Humanities (NEH) Office of Digital Humanities (ODH) Digging Into Data Challenge ("Digging by Debating"; PIs Allen, Börner, Ravenscroft, McAlister, Reed, and Bourget; award no. HJ-50092-12). The authors thank the Indiana University Cognitive Science Program for continued supplemental research funding, and especially for research fellowships for Jaimie Murdock and Robert Rose. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Just as Britain and America have been described as two nations separated by a common language, different academic disciplines often use the same words with divergent meanings [1]. Interdisciplinary research thus poses unique challenges for information retrieval (IR). Word sense disambiguation [2, 3], differing publication practices across disciplines [4–6] and disjoint authorship networks [7] pose special challenges to information retrieval for interdisciplinary work. When the dimension of time is added, terminological shifts [8, 9], changing citation standards [10–13], and shifting modes of scholarly communication [4, 5, 14, 15] all amplify the challenges for IR to serve the need of interdisciplinary scholars.

Widespread digitization of monographs and journals by HathiTrust [16, 17] and Google Books [18, 19] enable new longitudinal studies of change in language and discourse [8, 9, 12, 20–22], an approach known as "distant reading" [23]. These data-driven distant readings contrast with "close readings", in which short passages and particular details are emphasized for scholarly interpretation. Newly digitized materials, which enable distant reading, differ from born-digital scholarly editions in three key ways: First, the reliance on optical character recognition (OCR) over scanned page images introduces noise into the plain-text representations of the text. Second, the unstructured text does not contain any markup that may differentiate page header and footer information, section headings, or bibliographic information from the main text. Finally, metadata is often automatically extracted and lacks the provenance information important to many humanities scholars. Researchers seeking to marry these "distant readings" to more traditional "close readings" are impacted by these factors [24].

Our goal is to develop computational methods for scholarly analysis of large-scale digital collections that are robust across both the technological inconsistency of the digitized materials and the variations of meaning and practice among fields and across time. A further goal of our approach is that these methods should inform interdisciplinary research by suggesting novel interpretations and hypotheses. The methods should support scholars who wish to drill down from high level overviews of the available materials to specific pages and sentences that are relevant for understanding the various responses of scholars and scientists to contentious issues within their fields.

In this paper, we provide a case study that focuses on meeting these challenges within the interdisciplinary field of History and Philosophy of Science (HPS). HPS must not only bridge the humanities and the sciences, but also the temporal divide between historically-significant materials and the present [25–28]. We show how faceted search using a combination of traditional classification systems and mixed-membership models can go beyond keyword search to inform resource discovery, hypothesis formulation, and argument extraction in our test domain, delivering methods that can be generalized to other domains.

Using a novel approach of drill-down topic modeling—simultaneously reducing both the size of the corpus and the unit of analysis—we demonstrate how a set of 1,315 fulltext volumes obtained by a keyword search from the HathiTrust digital library is progressively reduced to six focal volumes that did not appear in the top ten results in the initial HathiTrust search. Topic modeling of these volumes at various levels, from whole book down to individual sentences, provides the contexts for word-sense disambiguation, is relatively robust in the face of OCR errors, and ultimately supports a system for semi-automatic identification of argument structure. We show how visualizations designed for macroanalysis of disciplinary scientific journals can be extended to highlight interdisciplinarity in arguments from book data [29]. This guides researchers to passages important for the kind of "close reading" that lies at the heart of scholarly work in the humanities, supporting and augmenting the interpretative work

that helps us understand the intellectual and societal contexts in which scientific writings are produced and received.

While the extension of computational methods such as these to various questions in the humanities may eventually provide ways to test specific hypotheses, the main focus of such research is likely to remain exploratory and interpretative, in keeping with the humanities themselves [24, 30]. This approach nevertheless shares something with the sciences: it is experimental to the extent that it opens up a space of investigation within which quantitatively defined parameters can be systematically varied and results compared. Such exploratory experimentation is common not just in the social sciences, but also in the natural sciences [31, 32].

Our study consisted of six stages. (1) We used a keyword search of the HathiTrust collection to generate an initial corpus and we used *probabilistic topic models* on these volumes. (2) We exploited the *mixed-membership* property of the topic models to identify the multiple contexts of the selected volumes and reduce the original search space even further. (3) Because topic models define the notion of a document flexibly, we drilled down further by constructing *page-level topic models* of the reduced set of volumes selected at the previous stage. (4) We used the page-level results to rank books and select pages from them for closer analysis, demonstrating an approach to semi-automatic *argument extraction* which showcases the interpretive results of our search process. (5) We exploited the close reading of arguments for exploratory investigation of drilling down even further, to *sentence-level topic modeling* within a single volume. (6) We used *scientific mapping* to locate relevant volumes [33]. Because current science maps represent journal data, and data overlays are created based on journal names, we needed to construct a *classification crosswalk* from the UCSD Map of Science to the Library of Congress Classifications of these journals, finally allowing us to project books onto the science map.

We assessed success in our case study in three ways: (1) by the effectiveness of the process in leading non-experts to drill down to highly-relevant content in a very large collection of books; (2) by the ability of this process to spotlight a somewhat forgotten woman scientist who is important to the history of psychology; (3) by the capacity of the process to lead domain experts to a surprising discovery about the breadth of species discussed in these historical materials, thus enriching the historical context for current discussions of intelligence in microscopic organisms [34, 35]. Our assessments are qualitative rather than quantitative in nature, but they are appropriate given current limitations in quantitative assessments of the quality of topic models [36–38].

Related work

The use of topic models for information retrieval is not itself novel, having prior general applications [39, 40], scientific applications [41, 42], and humanities applications [43]. Similar to our approach, some of these applications support finer-grained retrieval by remodeling a subset of the corpus. The key novelty of our approach, is that we simultaneously alter the granularity of the documents in our models as we go from modeling books in collections, to pages in books, to sentences in pages.

Previous studies indicate a general consensus that human judgments about what makes a "good" topic are generally convergent. However, human judgment does not typically correlate well with quantitative measures of model fit [36], suggesting that people are interpreting the topics using as-yet poorly understood semantic criteria. Furthermore, variation among people in their interpretation of topic quality may be dependent upon expertise. Some topics that are poorly-rated by non-experts may in fact be judged highly coherent by experts who understand why certain documents have high membership in the topic, in contrast to non-experts who

focus solely on the highest-probability terms in the topic without knowledge of the underlying corpus [38]. Interactive topic modeling [44] approaches this issue by introducing human-in-the-loop topic selection and biasing measures that increase human judgment of topic model fitness. Our drill-down topic modeling approach does not require human feedback during the modeling stage, but during the corpus selection phase. This reduces the training cost of our approach and makes it more accessible for exploratory search.

The use of visualization techniques in information retrieval is well documented: Doyle's "Semantic Road Maps for Literature Searchers" explicitly justified the use of visualization as a summary of scientific literature, in particular as a time-saving measure by quickly showing relevant features of a document [45]. Doyle also emphasizes that even if a visualization is itself static, it is the result of a dynamic process of iterative remodeling and learning from new data. The UCSD Map of Science is a basemap that needs to be learned—just like a geographic map of the world—but that can subsequently be used to quickly gain an overview of the topical distribution of documents [29]. Visualization of semantic models is also well-documented, especially for topic models [46–48]. Prior models, including the results of LSA, word co-occurrence, and other semantic analyses were also visualized (see [33] for a timeline). The last step of the workflow described in this paper uniquely projects a topic model analysis onto a visualization base layer derived from different data (journal citation links) for different purposes (visualizing the citation structure of current science). While the sorted lists we provide below are useful for determining what to read next, visualization helps users to understand patterns, trends, and outliers, supporting quick evaluation of which items are most relevant to their interests.

Materials

HathiTrust Digital Library

The HathiTrust Digital Library is a collaboration between over ninety institutions to provide common access and copyright management to books digitized through a combination of Google, Internet Archive, and local initiatives. As of October 24, 2016, it consisted of over 14.7 million volumes represented both as raw page images and OCR-processed text (https://www.hathitrust.org/statistics_info).

Due to copyright concerns, fulltext access to page images and their OCR-processed counterparts is given only to pre-1928 materials, which are assumed to be in the public domain in the United States. When the work described in this paper was initiated in 2012, the public domain portion of the HathiTrust consisted of approximately 300,000 volumes. At the end of the funding period in 2014, the public domain consisted of 2.1 million volumes. As of October 24, 2016, that number stood at 5.7 million volumes, and it has continued to grow since then. During the funding period for this project, even summary data describing the fulltext of post-1928 materials were impossible to access for computational analysis from the HathiTrust. Recently, however, the HathiTrust Research Center (HTRC) Data Capsule has been developed to enable tightly restricted access to features extracted from in-copyright materials [49].

While the corpus size has increased more than 20-fold, the methods presented in this paper are aimed to reduce the portion of the corpus for analysis. For example, the first step described below involves topic modeling the results of a *keyword search*, resulting in a corpus of 1,315 volumes (which we referred to as *HT*1315). Using the same query on October 24, 2016, we returned 3,497 volumes. Both of these datasets are computationally-tractable for topic modeling on modern workstations, in contrast (for example) to the 1.2 terabyte HTRC Extracted Features Dataset, derived from 4.8 million volumes [50]. The methods described in detail below further reduced the *HT*1315 corpus to a smaller corpus of 86 volumes (*HT*86) which we

modeled at the page level. This corpus was then further analyzed and refined to a 6-volume collection for argument mapping (*HT*6).

Stop lists

Before analyzing the texts, it is common to apply a 'stop list' to the results, which excludes words that are poor index terms [51]. Frequently, these are high-frequency words such as articles ('a', 'an', 'the'), prepositions ('by', 'of', 'on'), and pronouns ('he', 'she', 'him'), which contain little predictive power for statistical analysis of semantic content [52]. We use the English language stop list in the Natural Language Toolkit, which contains 153 words [53]. Additionally, we filtered words occurring five or fewer times, which both excludes uncommon words and infrequent non-words generated by OCR errors. We also developed custom methods for stripping headers and footers from the OCRed pages provided by the HathiTrust, cleaning up hyphenated words crossing lines and page breaks, and obtaining volume metadata. Our source code is freely available at https://github.com/inpho/vsm/blob/master/vsm/extensions/htrc.py.

UCSD map of science

For our macroanalysis, we want to see how our selected texts divide among the different academic disciplines. As a base map for the disciplinary space (analogous to a world map for geospatial space), we use the UCSD Map of Science [29] which was created by mining scientific and humanities journals indexed by Thomson Reuters' Web of Science and Elsevier's Scopus. The map represents 554 sub-disciplines—e.g., Contemporary Philosophy, Zoology, Earthquake Engineering—that are further aggregated into 13 core disciplines, appearing similar to continents on the map—e.g., Biology, Earth Sciences, Humanities. Each of the 554 sub-disciplines has a set of journals and keywords associated with it.

Library of Congress Classification Outline (LCCO)

The Library of Congress Classification Outline (LCCO) is a system for classifying books, journals, and other media in physical and digital libraries. It is different from the Library of Congress Control Number (LCCN), which provides an authority record for each volume. The HathiTrust stores the LCCN, which we then use to query the Library of Congress database for the call number, which contains the LCCO, providing us with a disciplinary classification for each volume in the *HT*1315, *HT*86, and *HT*6 datasets.

Target domain: History and philosophy of scientific work on animal cognition

Our specific test domain is the history and philosophy of scientific work on animal cognition [54–56]. We aimed to identify and extract arguments about anthropomorphism from a relevant subset of the scientific works published in the late 19th and early 20th century. This period represents a critical time for the development of comparative psychology, framed at one end by the work of Charles Darwin and at the other end by the rise of the behaviorist school of psychology (see [57] for a full historical review). Using the methods described in this paper, we progressively narrowed the 300,000 volumes to a subset of 1,315 selected for topic modeling at the full-volume level, then 86 of these selected for page-level topic modeling, and then 6 specific volumes selected for manual analysis of the arguments.

The term 'anthropomorphism' itself illustrates the problem of word sense disambiguation. In theological and anthropological contexts, 'anthropomorphism' refers to the attribution of human-like qualities to gods. In the animal cognition context, it refers to the projection of

human psychological properties to animals. Given the theological controversy evoked by Darwin, our inquiry demands our system be robust in partitioning these separate discourses.

Methods

Methods overview

We followed a six-stage process, summarized in Fig 1. Each step is described in more detail further below. We introduce them briefly here:

1. LDA Topic modeling of a subset of volumes from the HathiTrust Digital Library selected by a keyword search, treating each volume as the unit document for the LDA process.
2. Querying the model to further reduce the original set of documents to a more relevant subset for our HPS objectives.
3. Drill-down LDA topic modeling on the smaller set treating individual pages as the unit documents, using this page-level model to select pages for further analysis.
4. Mapping of arguments on the selected pages by manual analysis, supported by the enhanced Online Visualisation of Arguments (OVA+) tool [58].
5. LDA topic modeling of single books, treating each *sentence* as document unit.
6. Mapping identified volumes onto UCSD Map of Science via a crosswalk from Library of Congress classification data to the journals used to construct the basemap.

Detailed methods

1. From keyword search to probabilistic topic modeling. We reduced the number of volumes to be routed to the topic modeling process by conducting a keyword search in the

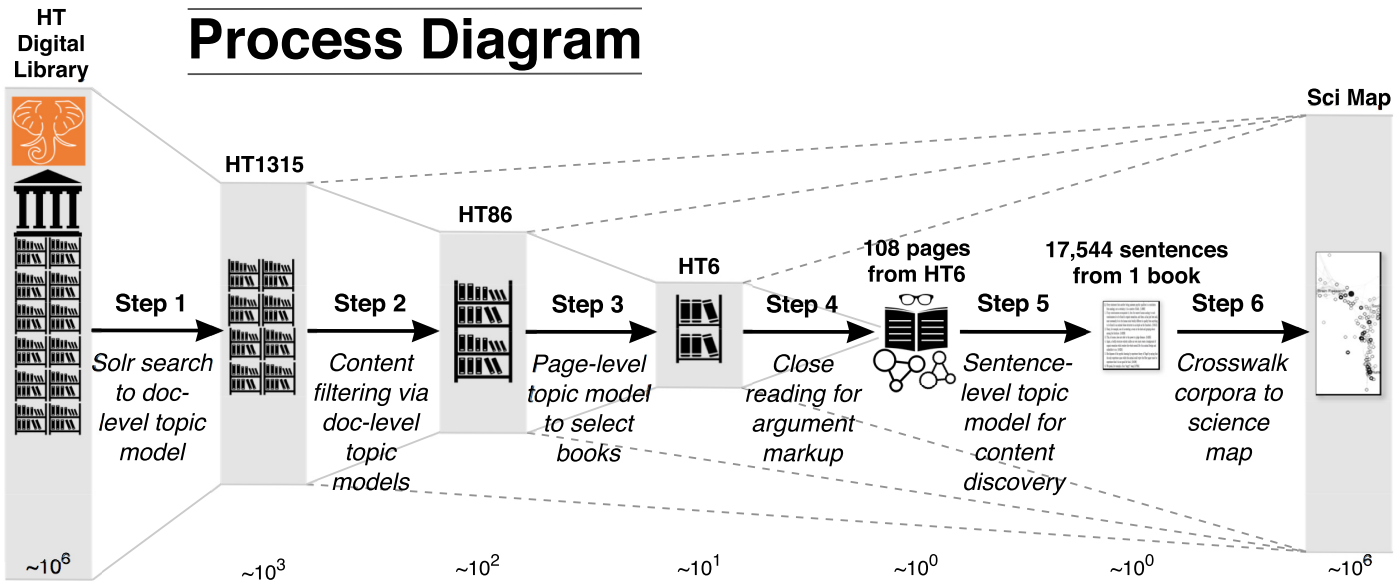

Fig 1. Corpus analysis sequence. Schematic rendering of the six-step process that sequentially drills down from macroscopic "distant reading" to microscopic "close reading" before zooming back out to the macroscopic scale at the final step. The approximate orders of magnitude of the datasets either side of each processing step are shown below the icons as powers of 10 of book/fulltext-sized units, and grey bars representing the data are scaled logarithmically.

https://doi.org/10.1371/journal.pone.0184188.g001

HathiTrust collection using the HathiTrust's Solr index. We searched using terms intended to reduce the hundreds of thousands of public domain works to a set of potentially relevant texts that could be efficiently modeled with the available computing resources. Specifically, we searched for "Darwin", "comparative psychology", "anthropomorphism", and "parsimony". While the specificity of our query may be seen as too restrictive, we emphasize (a) that we are following an exploratory research paradigm—we are not narrowing in on a particular fact, but rather surveying the available literature at the intersection of our interest in the history and philosophy of animal mind and cognition, and (b) the results of the keyword search were not specific enough to make the topic modeling redundant. Because we retrieved 1,315 volumes from the HathiTrust by this method, we refer to this corpus as *HT*1315. (More details can be found in the Results section below.)

Probabilistic topic models [37] are a family of mixed-membership models that describe documents as a distribution of topics, where each topic is itself a distribution over all words in a corpus. Topic models are *generative* models, that we interpret as providing a theory about context blending during the writing process [59].

Corpus preparation begins by treating each document as a bag of words. Common function words (prepositions, conjunctions, etc.) were filtered out using the NLTK stopword list for English, and rare words were filtered using a lower bound of 5 occurrences in the corpus. To construct the topic models used in this study, we use *Latent Dirichlet Allocation* (LDA— [60]) with priors estimated via Gibbs sampling [41] as implemented in the InPhO Topic Explorer [48].

The topic-modeling process begins by assigning random probabilities to the word-topic distributions and to each of the topic-document distributions. These prior distributions are then jointly sampled to generate estimates of the likelihood of observing the actual documents. These estimates are used to adjust the prior distributions in accordance with Bayes' rule. We ran this generate-and-test procedure iteratively for 1000 cycles, a number of iterations at which the distributions become relatively stable. Hyperparameters α and β control the word-topic and topic-distributions. We set them equal to 0.1, representing the expectation that each document should be weighted toward a mixture in which a relatively small subset of the available topics (k) dominate, and that topics should similarly be dominated by a relatively small proportion of the available words in the corpus. We initially modeled the *HT*1315 volumes using four different values for k, i.e., $k \in \{20, 40, 60, 80\}$.

2. Querying the models. At the end of the modeling process, each document is represented as a probability distribution over the k topics. We manually inspected the topics generated for the different values of k and determined that while all four of the models produced interpretable results, $k = 60$ provided the best balance between specificity and generality for our HPS goals.

We use the topic model to further narrow the search by querying topics with a combination of words. We do this by finding the topic or topics with the highest sum of the probabilities for each word. By a combination of trial and error, we found that a topic query combining 'anthropomorphism', 'animal', and 'psychology' produced more relevant topics and any term alone.

Using three topics identified in this way, we filtered the originally modeled set of books to a much smaller sub-corpus. The topic-document and word-topic distributions can be treated as vectors in their respective topic and word spaces. Thus it is possible to take the widely-used measure of vector cosines to assess similarity between topics and volumes. We computed the cosine distance between each of the three topics and the book's mixture of topics represented in the model. We summed these three distances and filtered them at the threshold of 1.25, yielding a smaller 86-book corpus (*HT*86) for more detailed analysis. The cutoff was chosen by

trial and error, manually inspecting the titles of the first few books excluded at a given threshold. Although more sophisticated selection methods exist (e.g., [39]) this approach was easy to understand and simple to implement by the team member tasked with identifying the arguments.

3. Drill down to page level. The notion of a "document" in LDA topic modeling is flexible. One can consider a full volume as a single document with a particular topic distribution. However, finer-grained models can also be made, in which each page, paragraph, or sentence receives its own topic distribution. Since OCR document scans in the HathiTrust have very little structural information—there is no encoding for section headings or paragraph breaks, let alone chapter breaks—the printed page was the next level below the full volume that we could reliably recover.

Hence, we re-modeled the *HT*86 set at the level of individual pages again using LDA topic modeling for values of $k \in \{20, 40, 60, 80\}$, parameterized as before, towards the goal of identifying arguments in text by "zooming in" to select books which had a high number of apparently relevant pages. For the sake of direct comparison to results reported above with the *HT*1315 model, we probed the $k = 60$ page-level model with 'anthropomorphism' as the query term alone, and in combination with other terms 'animal' and 'psychology' used previously. This identified one topic as most relevant to our project (see Results for details). We ranked volumes from the *HT*86 corpus according to which had the most pages among the top 800 highest ranked pages according to this topic and selected the top six volumes for the next step of the process (*HT*6). (The choice of six here was limited by time and resources allocated to the manual extraction of arguments detailed in the next section.)

4. Argument extraction: From pages to arguments. The selected pages were annotated using the Argument Interchange Format ontology (AIF [61]), which defines a vocabulary for describing arguments and argument networks. One of the coauthors [SM], who is not a domain expert, identified arguments using a semi-formal discourse analysis approach (informed by [62, 63]), and following a rubric established by the project PIs with HPS expertise [CA, DB, and AR]. The rubric supported identification of arguments based on their content and propositional structure, where this was also aided by noting argument signifiers in the texts, such as 'because', 'hence', 'therefore', etc. (Additional details about the rubric can be found in section 2.3.3 of [64].) This allowed us then to generate argument maps in the form of AIF annotated documents constructed with OVA+ (the enhanced Online Visualization of Arguments tool), available at http://ova.arg-tech.org/ (see also [58]). OVA+ provides a drag-and-drop interface for analyzing textual arguments, linking blocks of text as argument nodes. It also natively handles AIF structures. Each argument was divided into propositions and marked up as a set of text blocks. These text blocks containing propositions were linked to propositions that they support, or undercut, to create argument maps. OVA+ thus produces a visual representation of the structure of each argument.

5. Drilling down again: From arguments to sentences. To further investigate the utility of combining distant reading methods with close reading, we applied topic modeling to the sentences within a single volume. For this test we selected Margaret Washburn's *The Animal Mind* textbook [65] because it was top-ranked for topical content in *HT*6. We applied LDA topic modeling to its 17,544 sentences, treating this set of sentences as a collection of documents. To explore the power of topic modeling to identify latent but meaningful relationships at the micro-level, we arbitrarily chose a sentence from an Argument extracted from the Washburn set and used it to query the sentence-level model of *The Animal Mind* for the most similar sentences using the cosine of the sentence-topic vectors.

6. Zooming out: Macroanalysis by science mapping. At the final step, we created a visualization of the retrieved books overlaid on the UCSD Map of Science [33], to help understand the distribution of the retrieved books with respect to scientific disciplines.

In previous research, new datasets have been overlaid on the UCSD map by matching records via journal names or keywords to the 554 sub-disciplines. However, our present study is the first time that book data have been overlaid on a science map. To accomplish this, we constructed a *classification crosswalk* to align the journal-based sub-disciplines with a book classification system. The Library of Congress Classification Outline (LCCO) provides a hierarchical disciplinary taxonomy similar to that of the UCSD Map of Science. By using the Library of Congress Control Numbers (LCCN) assigned to each of the 25,258 journal sources in the UCSD Map of Science, we were able to use the hierarchical structure of the LCCO to assign a likelihood to any given book LCCN belonging to a particular USCD sub-discipline.

A number of items in the HathiTrust collection never received LCCNs. For example, university library collections frequently contain course bulletins that are not catalogued by the Library of Congress. We removed the uncatalogued items and projected the remaining volumes onto the UCSD map of science. We assigned each remaining book in *HT*1315 a UCSD sub-discipline based on its LCCN.

Results: A case study

In this section we describe the application of these methods to a case study in the History & Philosophy of Science (HPS), specifically in the history of comparative psychology. When we began the study the HathiTrust digital library provided access to the full texts of just over 300,000 public domain works. The keyword-based search for items of interest reduced this set to a corpus of 1,315 volumes published between 1800 and 1962, which we designate as our *HT*1315 corpus. (Publication dates after 1928 correspond to items in the public domain, such as government reports and university course bulletins.) A list of titles and HT handles is provided in the supplemental materials. Because the HT collection has changed over time, this exact set of results cannot be recreated by doing the same keyword search at hathitrust.org (see http://bit.ly/1LBbqnS). Currently there are over 5.5 million public domain works in the collection (see https://www.hathitrust.org/visualizations_dates_pd). The same query conducted in August 2015 yielded 3,027 full-text results.

Table 1 shows the top topics when the $k = 60$ topic model is queried using the single word 'anthropomorphism'. The topic model checking problem [37]—i.e., how to assess the quality of the model's topics—remains an important open problem in topic modeling. Nevertheless, most of the topics in the model can be quickly summarized. Inspection of this list indicates

Table 1. Topics ranked by similarity to 'anthropomorphism' in the *HT*1315 corpus. Topic 16 (highlighted with bold text) is highly relevant to the inquiry.

Topic	10 most probable words from topic
38	god, religion, life, man, religious, spirit, world, nature, spiritual, divine
16	**animals**, **evolution**, **life**, **animal**, **development**, **man**, **species**, **cells**, **living**, **theory**
51	philosophy, nature, knowledge, world, thought, idea, things, reason, truth, science
58	man, among, tribes, primitive, men, people, also, races, women, race
12	child, children, first, development, movements, play, life, little, mental, mother
21	social, life, new, mind, upon, individual, human, mental, world, subfield
11	motion, force, must, forces, matter, changes, us, parts, like, evolution
1	pp, der, vol, die, de, des, und, ibid, university, la
31	gods, religion, p, name, see, god, india, ancient, one, worship

https://doi.org/10.1371/journal.pone.0184188.t001

Table 2. Topics ranked by similarity to 'anthropomorphism', 'animal', and 'psychology' in the *HT*1315 corpus. Topics 26, 16, and 10 (highlighted with bold text) were used to derive the *HT*86 corpus, as they were most relevant to the inquiry.

Topic	10 most probable words from topic
26	**consciousness**, **experience**, **p**, **psychology**, **process**, **individual**, **object**, **activity**, **relation**, **feeling**
16	**animals**, **evolution**, **life**, **animal**, **development**, **man**, **species**, **cells**, **living**, **theory**
10	**animals**, **water**, **animal**, **food**, **birds**, **one**, **leaves**, **insects**, **species**, **many**
47	college, university, professor, school, law, work, students, degree, education, new
49	subfield, code, datafield, tag, ind2, ind1, b, d, c, controlfield
1	pp, der, vol, die, de, des, und, ibid, university, la
12	child, children, first, development, movements, play, life, little, mental, mother
58	man, among, tribes, primitive, men, people, also, races, women, race
21	social, life, new, mind, upon, individual, human, mental, world, subfield
2	test, tests, age, group, children, mental, table, per, cent, number

https://doi.org/10.1371/journal.pone.0184188.t002

that 'anthropomorphism' relates most strongly to a theological topic (38), a biological topic (16), a philosophical topic (51), an anthropological topic (58), and a child development topic (12). The topic model thus serves to disambiguate the different senses of 'anthropomorphism', especially between contexts where the discussion is about anthropomorphized deities (38) and contexts where it is about nonhuman animals (16), with the second topic being the most obvious attractor for researchers interested in comparative psychology. The second-to-last topic (1) is targeted on bibliographic citations, and is dominated by bibliographic abbreviations and some common German and French words that were not in the English language stop list used during initial corpus preparation. Although from one perspective this may seem like a 'junk' topic, this topic is nonetheless very useful to a scholar seeking citations buried in the unstructured pages in the corpus.

Table 2 shows the top topics returned by querying the $k = 60$ model of *HT*1315 using 'anthropomorphism', 'animal', and 'psychology' to construct the query. This new query reveals two relevant topics (numbers 26 and 10) that were not returned using 'anthropomorphism' alone. The top ten documents found by querying the model using these two topics in combination with the previously noted topic 16 is shown in Table 3. By selecting from the continuation of this list up to a threshold of 1.25 on the aggregated distance measure, we reduced the number of volumes of interest from 1,315 to 86, constituting the *HT*86 corpus.

Table 3. Book titles ranked by proximity of the full texts to topics 10, 16, and 26 in the *k* = 60 model of the *HT*1315 corpus.

Document	Distance
Secrets of animal life	0.87689
Comparative studies in the psychology of ants and of higher . . .	0.88814
The colours of animals, their meaning and use, especially . . .	0.98445
The foundations of normal and abnormal psychology	0.99833
The bird rookeries of the Tortugas	1.00286
Mind in animals	1.00294
Ants and some other insects; an inquiry into the psychic . . .	1.00504
Systematic science teaching: a manual of inductive . . .	1.01040
The riddle of the universe at the close of the 19th C.	1.01450
The colour-sense: its origin and development.	1.02795

https://doi.org/10.1371/journal.pone.0184188.t003

Table 4. Topics ranked by similarity to 'anthropomorphism' in the *HT*86 corpus, as modeled at the page level.

Topic	Top Ten Most Probable Words from Topic
18	god, religion, evolution, religious, man, human, science, world, christian, belief
3	mind, man, facts, life, evolution, instinct, subjective, instincts, organic, development
1	animal, animals, may, stimulus, experience, would, instinct, reaction, one, stimuli
51	sense, sensation, qualities, touch, perception, sensations, extension, sight, senses, us

https://doi.org/10.1371/journal.pone.0184188.t004

The result of querying the $k = 60$ page-level model of the *HT*86 corpus with the single query word 'anthropomorphism' is shown in Table 4. (Topic numbers are arbitrary and do not correlate across the *HT*86 and *HT*1315 models.) Although a theological topic (18) is again at the top of the list, it is clear that biological and psychological topics have become more prevalent in the *HT*86 model. Even within Topic 18, 'evolution' and 'science' are now among the ten highest probability words indicating that the topic is closer to a "religion and science" topic than the more general religion Topic 38 from the *HT*1315 model (Table 1), and reflecting the narrower range of books in the *HT*86 subset.

Using 'anthropomorphism', 'animal' and 'psychology' in combination to query the $k = 60$ *HT*86 model, topic 1 is the highest ranked topic (Table 5). In comparison to the earlier topics 10 and 16 from the *HT*1315 results in Table 2, this topic has more terms relevant to psychology (i.e., stimulus, experience, instinct, reaction), suggesting that for the purposes of locating specific pages in *HT*86 collection that are relevant to our HPS interests, topic 1 provides the best starting point. Table 6 shows the first rows of a list of 800 highest ranked pages from *HT*86 using topic 1 as the query.

None of the six volumes from *HT*86 collection which had the most pages in the top 800 highest-ranked pages had appeared among the top 10 keyword search results in the original Solr search in the HathiTrust collection. These volumes formed the *HT*6 collection:

1. *The Animal Mind: A Textbook of Comparative Psychology*, 1908 (first edition), by Margaret Floy Washburn, psychologist. Washburn's textbook was foundational for comparative psychology and she is notable as the second woman to be president of the American Psychological Association.
2. *Comparative studies in the psychology of ants and of higher animals*, 1905, a monograph by Erich Wasmann, an entomologist who only partly accepted evolution within species, rejecting common descent, speciation via natural selection, and human evolution.
3. *The Principles of Heredity*, 1906, a scientific monograph by G. Archdall Reid, a physician who argued against the Lamarckian idea of inheritance of acquired characteristics.
4. *General Biology*, 1910, a text book by James G. Needham, entomologist and limnologist.

Table 5. Topics ranked by similarity to 'anthropomorphism', 'animal', and 'psychology' in the *HT*86 corpus.

Topic	Top Ten Most Probable Words from Topic
1	animal, animals, may, stimulus, experience, would, instinct, reaction, one, stimuli
51	sense, sensation, qualities, touch, perception, sensations, extension, sight, senses, us
18	god, religion, evolution, religious, man, human, science, world, christian, belief
3	mind, man, facts, life, evolution, instinct, subjective, instincts, organic, development

https://doi.org/10.1371/journal.pone.0184188.t005

Table 6. Pages ranked by similarity to Topic 1.

Document	Distance
The animal mind, 1st ed., p. 43	0.04414
The animal mind, 2nd ed., p. 47	0.04552
The animal mind, 2nd ed., p. 263	0.10360
The animal mind, 2nd ed., p. 16	0.12336
The animal mind, 2nd ed., p. 71	0.15828
The animal mind, 1st ed., p. 219	0.16288
The animal mind, 1st ed., p. 232	0.16674
The animal mind, 1st ed., p. 57	0.18380
The animal mind, 1st ed., p. 72	0.22610
Mind in the lower animals, p. 179	0.23408

https://doi.org/10.1371/journal.pone.0184188.t006

5. *The Nature and Development of Animal Intelligence*, 1888, a compilation of articles by Wesley Mills, physiologist, physician and veterinarian.
6. *Progress of Science in the Century*, 1908, a book on the history of science for general readers by J. Arthur Thomson, naturalist.

These books were written by two Americans (Washburn and Needham), two Scots (Reid and Thomson), a Canadian (Mills), and an Austrian (Wasmann). They provide a broad array of perspectives on animal intelligence and psychology, from specialist monographs to textbooks to general-audience nonfiction, spanning both pro-Darwinian and anti-Darwinian viewpoints.

Using the $k = 60$ model of the *HT*6 collection to identify sections of each book with highest proportion of Topic 1, we selected 108 pages from the six *HT*6 volumes for further analysis (Table 7). From these we generated 43 argument maps using AIF annotated documents, providing a visual representation of the structure of each argument (e.g., Fig 2).

We performed two types of argument analysis: Pass 1 aimed to *summarize* the arguments presented in each volume. Pass A aimed to *sequence* the arguments presented in each volume. All argument maps can be found at http://bit.ly/1bwJwF9. A full description of the study, including analysis of the arguments can be found in [64], and is summarized in [66].

As a proof of concept, these arguments show the utility of new techniques for faceted search enabling access from a library of over 300,000 books to volume-level analysis of a subset of 1,315 books all the way down to page-level analyses of 108 pages for the purpose of identifying, encoding, modeling, and visualizing arguments. These argument diagrams function as a type of close reading, common in the humanities, where this approach is related to a range of work

Table 7. Pages for which OVA+ argument maps were created, showing total number of pages analyzed and numbers of arguments identified on each of the passes described in the main text.

Volume	Pages	Total	Pass 1	Pass A
The Animal Mind	13–16, 16–21, 24–27, 28–31, 31–34, 58–64, 204–207, 288–294	40	9	15
The Psychology of Ants	Preface, 15–19, 31–34, 48–53, 99–103, 108–112, 206–209, 209–214	37	8	10
The Principles of Heredity	374, 381, 382, 385, 386, 390, 394, 395	10		8
General Biology	434–435, 436	3		2
The Nature and Development of Animal Intelligence	16–18, 21–26, 30–32	12		5
Progress of Science	479–484	6		3
Overall Totals		108	17	43

https://doi.org/10.1371/journal.pone.0184188.t007

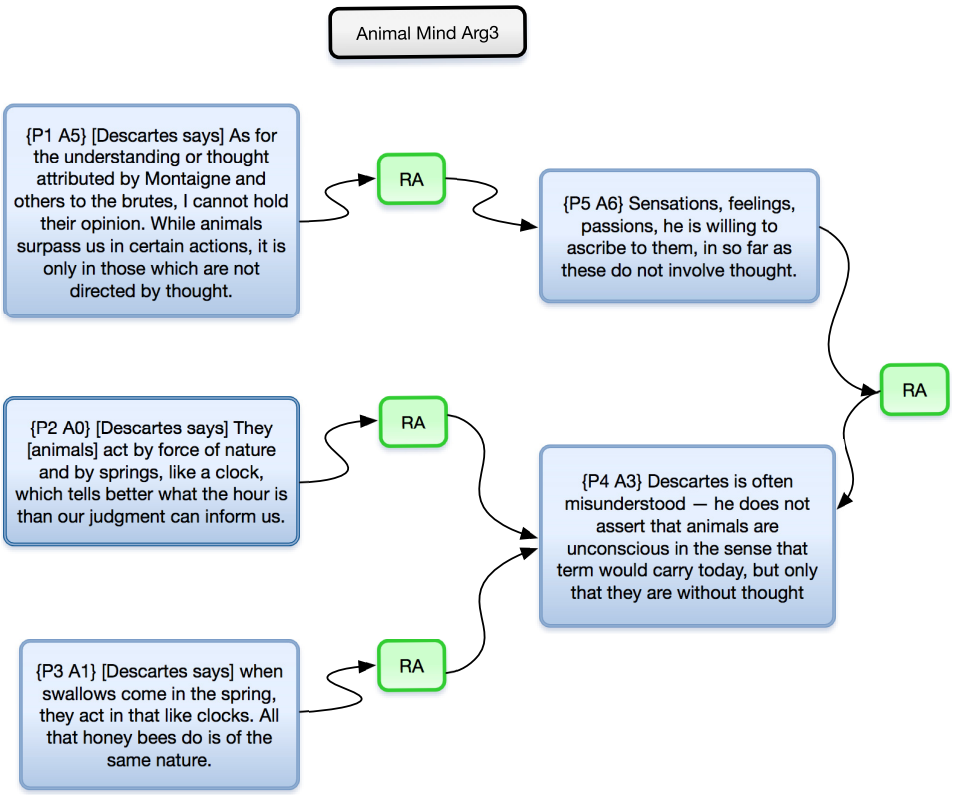

Fig 2. An argument map derived from *The Animal Mind*, represented in OVA+.

https://doi.org/10.1371/journal.pone.0184188.g002

in human-computer argument modeling [62, 67] and argument mapping Kirschner2012. This approach also draws on a rich tradition of philosophical literature (reviewed in [68]).

The query sentence (Q) that we chose from Argument 15 of the Washburn book is shown below with the first half dozen results (and their similarity scores). It is important to remember that LDA topic modeling is a "bag of words" approach; i.e., it uses only an unordered list of words in each document. It has no information about word order, punctuation, or other formatting in the text, and some of the most common words are not included. The full sentences are shown here only to aid the reader.

Q: Every statement that another being possesses psychic qualities is a conclusion from analogy, not a certainty; it is a matter of faith. (1.0000)

1. If any consciousness accompanies it, then the nearest human analogy to such consciousness is to be found in organic sensations, and these, as has just been said, must necessarily be in the human mind wholly different in quality from anything to be found in an animal whose structure is as simple as the Amoeba's. (0.8413)
2. Fancy, for example, one of us entering a room in the dark and groping about among the furniture. (0.8239)
3. This, of course, does not refer to the power to judge distance. (0.8235)
4. Again, a bodily structure entirely unlike our own must create a background of organic sensation which renders the whole mental life of an animal foreign and unfamiliar to us. (0.8224)

5. She disposes of the psychic learning by experience theory of Nagel by saying that the only experience upon which the animal could reject the filter paper must be experience that it is not good for food. (0.8198)

6. We speak, for example, of an "angry" wasp (0.7924)

Sentence 1 is obviously related in meaning to the query sentence: the sentences overlap in some words, and directly express related ideas. But the relevance of the other examples is less direct. Sentence 6 provides a nice illustration of anthropomorphic attribution with no word overlap whatsoever. The inclusion of sentences 2 and 3 is, more puzzling. However, in the context of where these sentences appear in Washburn's book, the relationship become plainer. Sentence 2 comes in the context of the discussion of what it might be like to be an amoeba. It is thus related to sentence 1, and it is used by Washburn to make the point that our experience in the dark, which still involves visual imagination and memories of what we touch, must be "wholly different in quality" (per sentence 1) from what an amoeba might experience. Sentence 3 occurs in a footnote on page 238, and it is worth quoting the footnote in full:

> Porter observed that the distance at which spiders of the genera Argiope and Epeira could apparently see objects was increased six or eight times if the spider was previously disturbed by shaking her web (612). This, of course, does not refer to the power to *judge* distance.
>
> [Italics in original.]

Here, then, we see the author cautioning the reader not to jump to a high-level interpretation of the spider behavior. The spiders may perceive objects at various distances but they don't judge it. The term 'judge' here is philosophically interesting, as it suggests an influence of Immanuel Kant on framing the debate. While Kant's name does not appear in Washburn's book, the term 'judgment' is important to Kant's theory of cognition, and fundamental to the cognitive divide he posits between humans and animals. We emphasize that this is just a speculative suggestion about Washburn's influences, but it does show how the topic modeling process can bring certain interpretive possibilities to the fore, moving the digital humanities another step closer to the goal of generating new insight into human intellectual activity.

Finally, we found Library of Congress classification records for 776 out of 1,315 books and we used the LCCO classification crosswalk that we constructed to locate these books on the UCSD Map of Science, as shown in Fig 3. (In the interactive online version, nodes can be selected, showing which volumes are mapped and providing the title and links to various external sources of metadata.) The map confirms that the initial keyword-based selection from the HathiTrust retrieved books that are generally positioned below the "equator" of the map, with particular concentrations in the life sciences and humanities, as was to be expected. The map provides additional visual confirmation that the further selections via topic modeling to the *HT*86 and *HT*6 corpora managed to target books in appropriate areas of interest.

Ultimately, the map overlay provides a grand overview and a potential guide to specific books that were topic modeled, although without further guidance from the topic models, the map does not fully meet the desired objective of linking a high-level overview to more detailed textual analysis.

General discussion

The notion of "distant reading" [23] has captured the imagination of many in the digital humanities. But the proper interpretation of large-scale quantitative models itself depends on having a feel for the texts, similar to Barbara McClintock's stress on having a "feeling for the

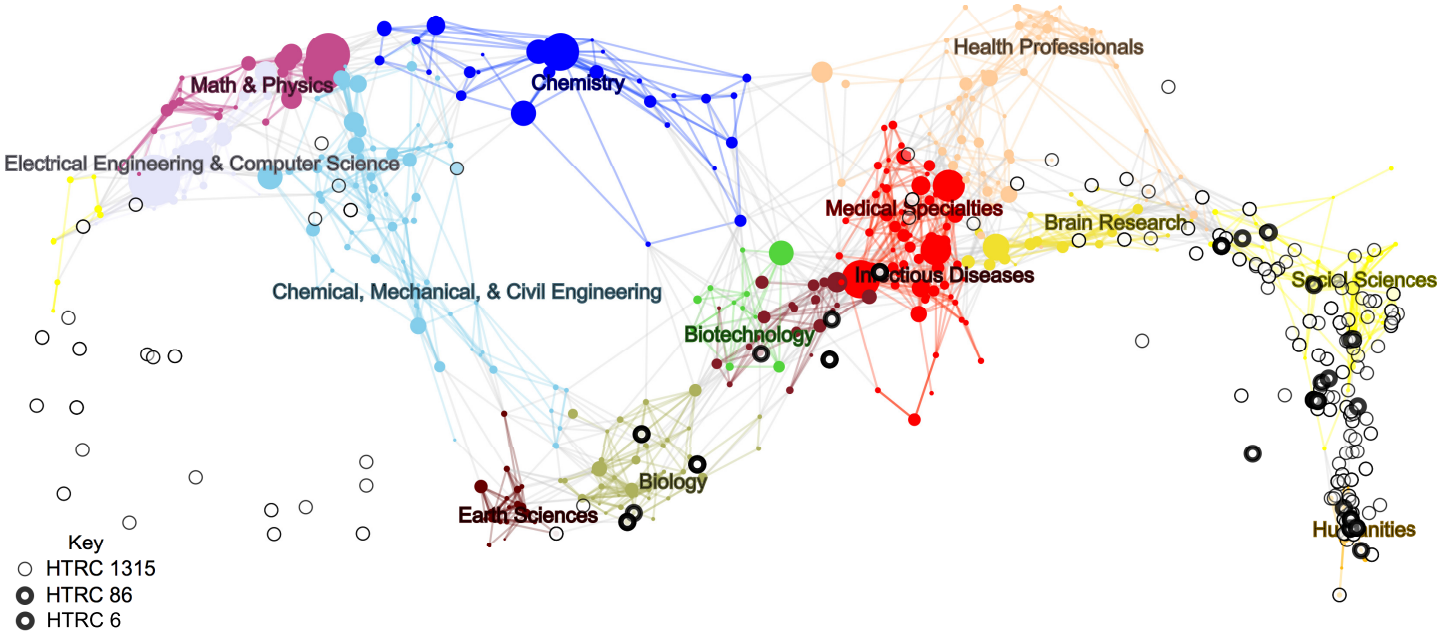

Fig 3. UCSD map of science with overlay of HathiTrust search results. This image shows topical coverage of humanities and life science data. The basemap of science shows each sub-discipline denoted by a circle colored or shaded according to the 13 core disciplines. Links indicate journal co-citations from the basemap. The 776 volumes of *HT*1315 with LCCN metadata are shown on the map as circles. Volumes also in *HT*86 are shown with thicker circles, and those in *HT*6 are shown in the thickest circles. An online, interactive version can be explored at http://inpho.cogs.indiana.edu/scimap/scits.

https://doi.org/10.1371/journal.pone.0184188.g003

organism" [69] or Richard Feynman on the importance for nascent physicists of developing "a 'feel' for the subject" beyond rote knowledge of the basic laws [70]. The interpretation of data and models, whether in science or the humanities, is itself (as yet, and despite a few small successes in fields such as medical diagnosis) a task at which humans vastly outperform machines. For this reason, the digital humanities remain a fundamentally hermeneutic enterprise [30], and one in which distant readings and close readings must be tightly linked if anything is to make sense.

In this paper we have motivated, introduced, and exemplified a multi-level computational process for connecting macro-analyses of massive amounts of documents to micro-level close reading and careful interpretation of specific passages within those documents. Thus we have demonstrated how existing computational methods can be combined in novel ways to go from a high-level representation of many documents to the discovery and analysis of specific arguments contained within documents.

We have also shown how to zoom out to a macro-level overview of the search results. We presented a novel classification crosswalk between the Library of Congress Classification Outline (LCCO) and the UCSD Map of Science, which was constructed using only journal data, to extend the data to books. Because of the mismatch between the book data and the journal metadata, the crosswalk is not perfect, and the method of averaging locations places many books in uninterpretable regions of the map. Nevertheless, the visualization provides some useful information about the effectiveness of a simple keyword search in locating items of interest within a collection of hundreds of thousands of books.

That our method succeeded in discovering texts relevant to a highly specific interdisciplinary inquiry shows its robustness to inconsistent and incomplete data. The HathiTrust Digital Library had OCR errors in 2.4% of volumes as of May 2010 [71]. While the quality of the

HathiTrust has increased in the intervening years, it is still a pervasive issue in digital archives [72].

Multi-level topic modeling combined with appropriate measures of distance can efficiently locate materials that are germane to a specific research project, going from more than a thousand books, to fewer than a hundred using book-level topic models, and further narrowing this set down to a small number of pages within a handful of books using page-level topic models. The similarity measures used to span the word-topic and topic-document distributions are mediated by the topics in the model, and because every topic assigns a probability to every word in the corpus, this approach is highly adept at finding implicit relationships among the documents. Typical applications of topic modeling used elsewhere, such as graphing the rise and fall of topics through time, may show large-scale trends, but do not directly facilitate the interplay between distant reading and close reading that leads to deeper understanding. By connecting abstract, machine-discovered topics to specific arguments within the text, we have shown how topic modeling can bridge this gap.

Could the similar outcomes have been accomplished with alternative methods such as Latent Semantic Analysis (LSA; [73]) or other Natural Language Processing (NLP) methods? We have investigated LSA as a discovery tool in the *HT*1315 and *HT*86 corpora. For users conditioned by online search engines, the word-centric search paradigm of LSA is more familiar than the topic-centric retrieval methods introduced here. Nonetheless, once users become aware of the potential for topic models, they provide more information about retrieved documents. For example, Washburn's *The Animal Mind* is significant for its mixture of topics. (The topic models of the *HT*1315 collection can be explored using Washburn's book as an entry point to the InPhO Topic Explorer page here: http://bit.ly/2qQZCJm.) Indeed the most prominent topic in the book (Topic 42: light, eye, two, eyes, visual, fig, vision, red, distance, movement) points to the importance of perceptual systems in her discussion of anthropomorphic attributions, and it is noteworthy that her dismissal of amoeba consciousness is grounded in what she argues is the lack of crossmodal associations (between, e.g., touch and vision) that are part of the content of human perceptual experiences—viz. the sentence previously discussed concerning why our human experience in the dark must be "wholly different in quality" from what an amoeba might experience.

Conclusions

Using a six-step process we progressively reduced 300,000 public domain volumes from the HathiTrust Digital Library to the 1,315 books in the *HT*1315 corpus, to the roughly 32,000 pages in the *HT*86 collection, to the over 17,000 sentences of the *HT*6 collection, to smaller set of the 108 pages selected for close reading and argument markup. This reduction allowed us to filter out discussions of anthropomorphic deities and zero in on key elements of late 19th and early 20th Century arguments about anthropomorphizing nonhuman organisms.

The application to the history of comparative psychology described in our case study was successful in at least three ways. First, team members unfamiliar with the domain of comparative psychology were effectively guided towards highly relevant material in an unreadably large set of books. Second, team members were introduced to the work of Margaret Washburn, a pioneer of scientific comparative psychology, who wrote an important textbook of the early 20th Century [65]—a book that went through four editions in as many decades, but has been largely forgotten since then (although see [74] for a recent tribute). Third, close reading of the arguments in the *HT*6 corpus revealed the surprising taxonomic range of these arguments, to include consideration even of consciousness in amoebae. This was surprising even to the domain experts on the team. This discovery raises new questions about the context for

contemporary discussions of slime mold intelligence [34, 35] and opens up new avenues for research and analysis. By putting words into context, topic modeling enabled researchers to zero in on passages worthy of detailed analysis and further humanistic interpretation.

It took Charles Darwin 23 years to read a number of books comparable in size to *HT*1315, as documented in his Reading Notebooks. At the unlikely rate of one book a day, it would take nearly four years to read this set of books in its entirety. Even allowing for the fact that one fifth of the volumes retrieved in our were course catalogs, eliminating those would nonetheless leave a daunting, if not quite Olympian, reading task. As the majority of the volumes selected by keyword search were not directly relevant to the research project, the payoff made possible by more sophisticated computational analysis of the full texts was critical for the present task of finding arguments about anthropomorphizing animals.

Although we do not claim that our methods provide the only possible approach, we are unaware of any other approach providing the kind of multilevel macro-to-micro approach needed to assist digital humanities researchers who wish to leverage computational methods against large data collections to support both distant reading and more traditional close reading analyses in the humanities. Nevertheless, these methods suffer from the problem already noted by Blei [37] and others that objective methods for evaluating the quality of topics are lacking. We regard this as a function of the complexity of human semantic understanding—the holy grail of strong Artificial Intelligence. And while studies such as ours raise more questions than they answer, they point the way to even deeper understanding of large text corpora and more useful tools to scholars.

Availability of data and materials

Materials

UCSD map of science. The UCSD Map of Science is available at http://sci.cns.iu.edu/ucsdmap/

LOC-UCSD crosswalk. The LoC-UCSD Crosswalk is available via GitHub at https://github.com/inpho/loc-ucsd.

Corpus and models. A tar.gz file containing the HTRC-1315, HTRC-86 corpora and trained models supporting the conclusions of this article is available via the Indiana University ScholarWorks at https://hdl.handle.net/2022/21636 (doi: 10.5967/K8251GBZ). The models comprise of:

1. *HTRC-1315, volume-level, 60 topics*
2. *HTRC-86, page-level, 60 topics*

All materials in each corpus are in the public domain. All models are released under a Creative Commons Attribution 4.0 International (CC BY 4.0) License.

Software. The InPhO Topic Explorer and `vsm` semantic modeling library are available on GitHub at https://github.com/inpho/topic-explorer and https://github.com/inpho/vsm, respectively. Both are published under the MIT License, a maximally-permissive open-source license.

The notebooks used for the HT1315 and HT86 analyses are included in the Indiana University Scholar Works at https://hdl.handle.net/2022/21636. A read-only version of the notebooks, without the underlying models, may be viewed at https://github.com/inpho/digging. All notebooks are released under a Creative Commons Attribution 4.0 International (CC BY 4.0) License.

Acknowledgments

This work was funded by the National Endowment for Humanities (NEH) Office of Digital Humanities (ODH) Digging Into Data Challenge ("Digging by Debating"; PIs Allen, Börner, Ravenscroft, Reed, and Bourget; award no. HJ-50092-12). The authors thank the Indiana University Cognitive Science Program for continued supplemental research funding, and especially for research fellowships for Jaimie Murdock and Robert Rose. We also thank the HathiTrust Research Center (HTRC) for their support of research activities and generous access to materials.

Author Contributions

Conceptualization: CA KB AR CR DB.

Data curation: RL JM RR DR JO SM JL.

Formal analysis: JM CA KB RL SM AL CR JL.

Funding acquisition: CA KB AR CR DB.

Investigation: JM CA RL KB SM AR JL CR DB.

Methodology: JM CA KB AR RR DB CR.

Project administration: SM CA AR KB.

Resources: CA KB AR CR.

Software: JM RL RR DR JO JL CR.

Supervision: CA KB AR DB CR.

Validation: CA KB AR.

Visualization: JM RL CA.

Writing – original draft: JM.

Writing – review & editing: JM CA KB RL SM AR.

References

1. Demarest B, Sugimoto CR. Argue, observe, assess: Measuring disciplinary identities and differences through socio-epistemic discourse. Journal of the Association for Information Science and Technology. 2015; 66(7):1374–1387. https://doi.org/10.1002/asi.23271
2. Mihalcea R. In: Sammut C, Webb GI, editors. Word Sense Disambiguation. Boston, MA: Springer US; 2010. p. 1027–1030.
3. Agirre E, Edmonds P. Word Sense Disambiguation: Algorithms and Applications. Text Speech and Language Technology. 2006; 33:384.
4. Borgman CL, Furner J. Scholarly communication and bibliometrics. Annual Review of Information Science and Technology. 2002; 36(1):2–72. https://doi.org/10.1002/aris.1440360102
5. Cronin B. Scholarly communication and epistemic cultures. New Review of Academic Librarianship. 2003; 9(1):1–24. https://doi.org/10.1080/13614530410001692004
6. Holmberg K, Thelwall M. Disciplinary differences in Twitter scholarly communication. Scientometrics. 2014; 101(2):1027–1042. https://doi.org/10.1007/s11192-014-1229-3
7. Kärki R. Searching for bridges between disciplines: an author co-citation analysis on the research into scholarly communication. Journal of Information Science. 1996; 22(5):323–334. https://doi.org/10.1177/016555159602200501
8. Hamilton WL, Leskovec J, Jurafsky D. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. Association for Computational Linguistics (ACL). 2016.

9. Hamilton WL, Leskovec J, Jurafsky D. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. CoRR. 2016.
10. Kaplan N. The norms of citation behavior: Prolegomena to the footnote. American Documentation. 1965; 16(3):179–184. https://doi.org/10.1002/asi.5090160305
11. Liu M. Progress in Documentation—The Complexities of Citation Practice: A Review of Citation Studies. Journal of Documentation. 1993; 49(4):370–408. https://doi.org/10.1108/eb026920
12. Larivière V, Archambault É, Gingras Y. Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). Journal of the American Society for Information Science and Technology. 2008; 59(2):288–296. https://doi.org/10.1002/asi.20744
13. de Rijcke S, Wouters PF, Rushforth AD, Franssen TP, Hammarfelt B. Evaluation practices and effects of indicator use—a literature review. Research Evaluation. 2016; 25(2):161–169. https://doi.org/10.1093/reseval/rvv038
14. Odlyzko A. The rapid evolution of scholarly communication. Learned Publishing. 2002; 15(1):7–19. https://doi.org/10.1087/095315102753303634
15. Evans JA. Electronic Publication and the Narrowing of Science and Scholarship. Science. 2008; 321(5887):395–399. https://doi.org/10.1126/science.1150473 PMID: 18635800
16. York J. This library never forgets: Preservation, cooperation, and the making of HathiTrust Digital Library. Archiving 2009: Final Program & Proceedings. 2009; 2009(1):5–10.
17. Christenson H. HathiTrust: A research library at web scale. Library Resources and Technical Services. 2011; 55(2):93–102. https://doi.org/10.5860/lrts.55n2.93
18. Coyle K. Mass Digitization of Books. The Journal of Academic Librarianship. 2006; 32(6):641–645. https://doi.org/10.1016/j.acalib.2006.08.002
19. Vincent L. Google Book Search: Document Understanding on a Massive Scale. In: International Conference on Document Analysis and Recognition, ICDAR'2007; 2007. p. 819–823.
20. Altmann EG, Pierrehumbert JB, Motter AE. Beyond Word Frequency: Bursts, Lulls, and Scaling in the Temporal Distributions of Words. PLOS ONE. 2009; 4(11):e7678. https://doi.org/10.1371/journal.pone.0007678 PMID: 19907645
21. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, Team TGB, et al. Quantitative Analysis of Culture Using Millions of Digitized Books. Science. 2011; 331(6014):176–182. https://doi.org/10.1126/science.1199644 PMID: 21163965
22. Cocho G, Flores J, Gershenson C, Pineda C, Sánchez S. Rank Diversity of Languages: Generic Behavior in Computational Linguistics. PLOS ONE. 2015; 10(4):e0121898. https://doi.org/10.1371/journal.pone.0121898 PMID: 25849150
23. Moretti F. Distant Reading. London: Verso Books; 2013.
24. Underwood T. Theorizing Research Practices We Forgot to Theorize Twenty Years Ago. Representations. 2014; 127(1):64–72. https://doi.org/10.1525/rep.2014.127.1.64
25. Kuhn T. The Relations Between the History and the Philosophy of Science. In: The Essential Tension: Selected Studies in Scientific Tradition and Change. Chicago, IL: University of Chicago Press; 1979. p. 3–20.
26. Laudan L, Donovan A, Laudan R, Barker P, Brown H, Leplin J, et al. Scientific change: Philosophical models and historical research. Synthese. 1986; 69(2):141–223. https://doi.org/10.1007/BF00413981
27. Hacking I. Two Kinds of "New Historicism" for Philosophers. New Literary History. 1990; 21(2):343–364. https://doi.org/10.2307/469257
28. Weingart SB. Finding the History and Philosophy of Science. Erkenntnis. 2015; 80(1):201–213. https://doi.org/10.1007/s10670-014-9621-1
29. Börner K, Klavans R, Patek M, Zoss AM, Biberstine JR, Light RP, et al. Design and Update of a Classification System: The UCSD Map of Science. PLOS ONE. 2012; 7(7):1–10.
30. Rockwell G, Sinclair S. Hermeneutica. Cambridge, MA: MIT Press; 2016.
31. Steinle F. Entering New Fields: Exploratory Uses of Experimentation. Philosophy of Science. 1997; 64(S1):S65–S74. https://doi.org/10.1086/392587
32. Waters CK. The Nature and Context of Exploratory Experimentation: An Introduction to Three Case Studies of Exploratory Research. History and Philosophy of the Life Sciences. 2007; 29(3):275–284. PMID: 18822658
33. Börner K. Atlas of Science: Visualizing What We Know. Cambridge, MA: MIT Press; 2010.
34. Nakagaki T, Yamada H, Tóth A. Maze-solving by an amoeboid organism. Nature. 2000; 407:470. https://doi.org/10.1038/35035159 PMID: 11028990

35. Reid CR, MacDonald H, Mann RP, Marshall JAR, Latty T, Garnier S. Decision-making without a brain: how an amoeboid organism solves the two-armed bandit. Journal of The Royal Society Interface. 2016; 13:e1–e8. https://doi.org/10.1098/rsif.2016.0030

36. Chang J, Gerrish S, Wang C, Boyd-graber JL, Blei DM. Reading tea leaves: How humans interpret topic models. In: Advances in neural information processing systems; 2009. p. 288–296.

37. Blei DM. Probabilistic Topic Models. Communications of the ACM. 2012; 55(4):77–84. https://doi.org/10.1145/2133806.2133826

38. Lee TY, Smith A, Seppi K, Elmqvist N, Boyd-Graber J, Findlater L. The human touch: How non-expert users perceive, interpret, and fix topic models. International Journal of Human-Computer Studies. 2017; 105:28–42. https://doi.org/10.1016/j.ijhcs.2017.03.007

39. Wei X, Croft WB. LDA-based Document Models for Ad-hoc Retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'06. New York, NY, USA: ACM; 2006. p. 178–185.

40. Medlar A, Glowacka D. Using Topic Models to Assess Document Relevance in Exploratory Search User Studies. In: Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval. CHIIR'17. New York, NY, USA: ACM; 2017. p. 313–316.

41. Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences. 2004; 101(suppl 1):5228–5235. https://doi.org/10.1073/pnas.0307752101

42. Hall D, Jurafsky D, Manning CD. Studying the History of Ideas Using Topic Models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP'08. Stroudsburg, PA, USA: Association for Computational Linguistics; 2008. p. 363–371. Available from: http://dl.acm.org/citation.cfm?id=1613715.1613763.

43. Tangherlini TR, Leonard P. Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research. Poetics. 2013; 41(6):725–749. https://doi.org/10.1016/j.poetic.2013.08.002

44. Hu Y, Boyd-Graber J, Satinoff B, Smith A. Interactive topic modeling. Machine Learning. 2014; 95(3): 423–469. https://doi.org/10.1007/s10994-013-5413-0

45. Doyle LB. Semantic Road Maps for Literature Searchers. Journal of the Association for Computing Machinery. 1961; 8(4):553–578. https://doi.org/10.1145/321088.321095

46. Chaney AJB, Blei DM. Visualizing Topic Models. In: International AAAI Conference on Social Media and Weblogs; 2012.

47. Chuang J, Roberts ME, Stewart BM, Weiss R, Tingley D, Grimmer J, et al. TopicCheck: Interactive Alignment for Assessing Topic Model Stability. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics; 2015. p. 175–184. Available from: http://www.aclweb.org/anthology/N15-1018.

48. Murdock J, Allen C. Visualization Techniques for Topic Model Checking. In: Proceedings of the 29th AAAI Conference (AAAI-15). Austin, TX: AAAI Press; 2015. Available from: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10007.

49. Zeng J, Ruan G, Crowell A, Prakash A, Plale B. Cloud Computing Data Capsules for Non-Consumptive Use of Texts. In: ScienceCloud'14: Proceedings of the 5th ACM Workshop on Scientific Cloud Computing. Vancouver, BC, Canada; 2014. p. 9–16.

50. Capitanu B, Underwood T, Organisciak P, Bhattacharyya S, Auvil L, Fallaw C, et al. Extracted Feature Dataset from 4.8 Million HathiTrust Digital Library Public Domain Volumes (0.2)[Dataset]; 2015.

51. Fox C. A Stop List for General Text. SIGIR Forum. 1989; 24(1–2):19–21. https://doi.org/10.1145/378881.378888

52. Luhn HP. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development. 1957; 1(4):309–317. https://doi.org/10.1147/rd.14.0309

53. Bird S, Loper E, Klein E. Natural Language Processing with Python. O'Reilly Media Inc.; 2009.

54. Allen C, Bekoff M. Species of Mind: The Philosophy and Biology of Cognitive Ethology. Cambridge, MA: MIT Press; 1999.

55. Andrews K. Animal Cognition. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. summer 2016 ed. Metaphysics Research Lab, Stanford University; 2016.

56. Allen C, Trestman M. Animal Consciousness. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. summer 2015 ed. Metaphysics Research Lab, Stanford University; 2015.

57. Boakes R. From Darwin to Behaviourism: Psychology and the Minds of Animals. Cambridge University Press; 1984.

58. Lawrence J, Bex F, Reed C, Snaith M. AIFdb: Infrastructure for the Argument Web. In: Frontiers in Artificial Intelligence and Applications. vol. 245. Vienna: IOS Press; 2012. p. 515–516.

59. Murdock J, Allen C, DeDeo S. Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. Cognition. 2017; 159:117–126. https://doi.org/10.1016/j.cognition.2016.11.012 PMID: 27939837

60. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003; 3(4–5):993–1022.

61. Chesñevar C, Modgil S, Rahwan I, Reed C, Simari G, South M, et al. Towards an argument interchange format. The Knowledge Engineering Review. 2006; 21(4):293–316. https://doi.org/10.1017/S0269888906001044

62. Ravenscroft A, Pilkington RM. Investigation by Design: Developing Dialogue Models to Support Reasoning and Conceptual Change. International Journal of Artificial Intelligence in Education: Special Issue on Analysing Educational Dialogue Interaction: From Analysis to Models that Support Learning. 2000; 11(1):273–298.

63. Pilkington RM. Discourse, Dialogue and Technology Enhanced Learning. Abindgon, Oxfordshire and New York: Routledge; 2016.

64. McAlister S, Allen C, Ravenscroft A, Reed C, Bourget D, Lawrence J, et al. From Big Data to Argument Analysis and Automated Extraction—Final Report. Digging into Data, Round 2; 2014. Available from: http://diggingbydebating.org/wp-content/uploads/2014/04/DiggingbyDebating-FinalReport2.pdf.

65. Washburn MF. The Animal Mind: A Text-book of Comparative Psychology. 1st ed. New York: Macmillan; 1908.

66. Lawrence J, Reed C, Allen C, McAlister S, Ravenscroft A. Mining Arguments From 19th Century Philosophical Texts Using Topic Based Modelling. In: Proceedings of the First Workshop on Argumentation Mining. Baltimore, Maryland: Association for Computational Linguistics; 2014. p. 79–87. Available from: http://www.aclweb.org/anthology/W/W14/W14-2111.pdf.

67. Yuan T, Moore D, Reed C, Ravenscroft A, Maudet N. Review: Informal Logic Dialogue Games in Human-computer Dialogue. Knowl Eng Rev. 2011; 26(2):159–174. https://doi.org/10.1017/S026988891100004X

68. Reed C, Walton D, Macagno F. Argument diagramming in logic, law and artificial intelligence. The Knowledge Engineering Review. 2007; 22(1):87–109. https://doi.org/10.1017/S0269888907001051

69. Keller EF. A Feeling for the Organism. New York: W. H. Freeman and Company; 1983.

70. Feynman R. Feynmann Lectures on Physics. Redwood City, CA: Addison-Wesley; 1964.

71. Conway P. Measuring Content Quality in a Preservation Repository: HathiTrust and Large-Scale Book Digitization. In: Proceedings of 7th International Conference on Preservation of Digital Objects, iPres 2010. Vienna; 2010. p. 95–102. Available from: http://hdl.handle.net/2027.42/85227.

72. Kichuk D. Loose, Falling Characters and Sentences: The Persistence of the OCR Problem in Digital Repository E-Books. Libraries and the Academy. 2015; 15(1):59–91. https://doi.org/10.1353/pla.2015.0005

73. Landauer TK, Dumais ST. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. Psychological Review. 1997; 104:211–240.

74. Washburn DA. The Animal Mind at 100. The Psychological Record. 2010; 60. https://doi.org/10.1007/BF03395714